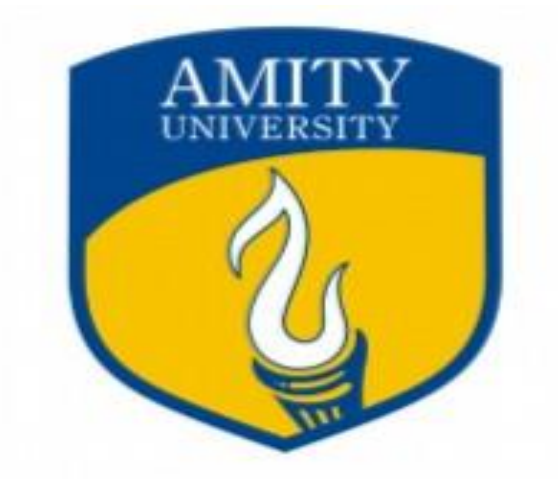


AMITY UNIVERSITY MADHYA PRADESH



ASSIGNMENT - I

of

TEXT & SOCIAL MEDIA ANALYTICS

CSD 501

SUBMITTED TO:-

DR. HEMANT KUMAR SONI

ASSISTANT PROFESSOR

SUBMITTED BY:-

KHUSHI PURWAR

B.Tech [CSE] - 5th Sem

Name - Khushi Purwan
Enrollment No. - A60205219013

Assignment - 1

Q1) How social media data influence decision making?
Explain data sources in social media?

A1) Social media are increasing their penetration into everyday life because of enhanced technology. These online technology tools help people use the Internet to communicate with friends and to share information and resources with our network of contacts. This results into the evolution of large amount of data.

Social media's impact on personal and managerial decision making can be extensive. Social media is altering our opinions and influencing our choices by impacting decisions of consumers and business decisions of managers. Social media manipulates the opinions.

While differences in the level of influence were found across industries, social media was deemed influential in making decisions and seeking advice. Social media was identified as being influential by 40% of respondents across different categories in their decision-making.

(2)

When making a decision, most consumers consult a social media channel. A study found that they found on social media influenced a massive 75%.

Overall, the impact of social media on decision making should increase in the future. Technologies are likely to increase the rationality and effectiveness of decision making in some cases while negatively impacting decision-making in other cases.

Social data sources used to be the information you gathered from the numerous social media channels.

Some of the biggest social media data sources are:

Facebook, Twitter, Instagram, YouTube, Blogs, etc.

XXXX

Q2) What is Tokenization? Explain its uses in social media analytics.

A2) Tokenization is the process of breaking sequence of strings into piece such as words, keywords, phrases, symbols and other elements called tokens.

Tokenization is the most important step in text processing. Different tools are available to perform the tokenization process.

a) Sentence Tokenization

In this tokenization, splitting the text into sentences.

b) Word Tokenization

In this, text splits into words.

c) Punctuation-based Tokenizer

This tokenizer splits the sentences into words based on whitespace and punctuations.

These are the most used tokenizer in text processing.

④

Q3) What is supervised learning? How its uses in social media analytics?

A) Supervised learning is the type of machine learning in which machines are trained using well labelled training data on the basis of that, machine produces output. Here, labelled data means some input data is already tagged with the correct output.

Supervised learning, the training data provided to the machine work as the supervisor that teaches the machine to predict the output correctly. The aim of supervised learning algorithm is to find a mapping function to map the input variable with output variable.

There are two types of supervised learning,

1) Regression: These algorithms are used if there is a relationship between the input variable and the output variable. For example, linear regression, etc.

2) Classification: These algorithms are used to when output variable is categorical, means there are two classes. For example, Random forest, decision trees, etc.

Social media is a massive landscape for communication that has a barrage of information about your brand. The key is to find the relevant and adds to our business. This is where machine learning plays a role.

The entire social network can be analyzed through machine learning for interpreting messages that represent customer satisfaction, anger, happiness and distress. We can learn more about the customer sentiments by implementing machine learning in our social media monitoring process.

The main objective of social media monitoring is to improve the decision-making process. Machine learning enhances social media monitoring to provide better insights and more detailed information.

XXXX

(6)

Q4, What is feature selection? Explain information gain.

A4, Feature selection is the process of selecting the subset of the relevant features and leaving out the irrelevant features present in a dataset to build a model of high accuracy. In simple words, it is a way of selecting the optimal features from the input dataset.

Three methods are used for feature selection.

1 Filters method

In this method, the dataset is filtered and a subset that contains only the relevant features is taken. Some common techniques are: Correlation, Chi-square test, information gain, etc.

2 Wrappers Method

It takes a machine learning model for evaluation. In this method, some features are fed to the ML model and evaluate the performance.

3 Embedded Methods

This method checks the different training iterations of the machine learning model and evaluate the importance of each feature.

(4)

Information gain is one of the technique for Attribute Selection Measure (ASM). It is the measurement of changes in entropy after the segmentation of a dataset based on a attribute. It calculate how much information a feature provide us about a class.

$$\text{Information gain} = \text{Entropy}(S) - [(\text{weighted avg}) * \text{Entropy}(\text{each feature})]$$

where Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as :

$$\text{Entropy}(S) = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 P(\text{No})$$

where, S = Total no. of samples

$P(\text{yes})$ = probability of yes

$P(\text{No})$ = probability of no.

XXX

(8)

Q5] How we can Extracting Social Media Data using API'S ?

Ans] Process of extracting data from social networks is called social media data scraping.

API (Application Programming Interface) is a set of methods of communicating between various software components. A API is used to send the request to a web server and in return server provides reviews or comments in a raw data format.

While extracting data from social media, a piece of code is running is called scraper. As it runs, get query rolls out to extract the HTML data coming from API on social media channels.

Thereafter, algorithms analyze a string of symbols, either in natural language or model is Document Object Model (DOM) structure. This parsing process determines nodes (an object represents a part of the document). Then it creates a nodes processor to show output in a normalized format.

In simple words, scraper comes into play, filtering through data to pick up the requisite datasets. Once the requirements is fulfilled, the data is translated into specific format.

XXX